Research in Natural Language Processing
January 15, 1985 - September 15, 1987

Ralph Grishman
PROTEUS Project Memorandum #12
November 1987

Research in Natural Language Processing
January 15, 1985 - September 15, 1987

This report describes research done by the PROTEUS Project at New York University during the period January 15, 1985 to September 15, 1987. All of the activities described below were supported in part by the Strategic Computing Program of the Defense Advanced Research Projects Agency under Contract N00014-85-K-0163 from the Office of Naval Research. Some of these activities were also supported by the National Science Foundation under Grant DCR-8501843. Development of the core syntactic analyzer was also supported by the Naval Research Laboratory under Contract N00014-85-K-2028 from the Office of Naval Research.

This report is intended to provide only an overview and outline of the various research activities performed during this period. Technical details of the various activities are provided by the Proteus Project Memoranda cited herein. In particular, a current and detailed presentation of the CASREP Analysis System which has been the primary product of the contract period is provided by Memorandum #11 (see section 2.3).

1. Core Syntactic Analyzer

1.1. Objective

The PROTEUS Syntactic Analyzer is intended to provide an efficient, easy-to-use base for the various experiments in computational linguistics described below.

1.2. Accomplishments

A syntactic analyzer has been built based on the chart parsing algorithm and using a compositional technique based on lambda reduction for syntactic regularization. A PROTEUS Restriction Language compiler is provided which translates a declarative language suitable for stating grammatical constraints into LISP code to check these constraints during parsing.

The analyzer was originally coded in Franz LISP, then ported to Symbolics Zetalisp and most recently to Common Lisp (running both on Symbolics LISP machines and on the SUN under Franz Extended Common LISP). The program has been distributed to the Naval Research Laboratory under a parser development contract with the laboratory, and has also been provided to Monmouth College, New Jersey as part of our budding research in machine translation.

1.3. Documentation

Proteus Project Memorandum #4: PROTEUS Parser Reference Manual. R. Grishman, July 1986

Proteus Project Memorandum #5: Syntactic Regularization in Proteus. J. M. Gawron, September 1986

Proteus Project Memorandum #9: An Introduction to the PROTEUS Parser. R. Grishman, March 1987

2. CASREP Analysis

2.1. Objective

Our basic, long-term objective, as part of the Strategic Computing Program in Natural Language Processing, is to develop the technology necessary for the robust automated processing of messages containing natural language narrative. One aspect of the development of such language processing systems is the incorporation of detailed domain knowledge and the effective use of such knowledge in language analysis. We have therefore focussed our research during this period on one type of message, CASREPs (equipment casualty reports), on developing detailed domain knowledge (a model of the equipment), and on using this knowledge for language understanding.

2.2. Accomplishments

We have selected one large piece of equipment, the starting air system for propulsion gas turbines, which is of substantial yet manageable complexity and has had frequent reported failures. We constructed a detailed model of this equipment, incorporating sufficient structural and functional information to permit a qualitative simulation of the system. Based on a corpus of 36 CASREPs involving this equipment, we have developed a CASREP grammar (using the PROTEUS Syntactic Analyzer mentioned above). In cooperation with UNISYS Defense Systems we developed a semantic analysis procedure. Finally, in the winter and spring of 1987 we developed an initial implementation of a discourse analyzer capable of identifying implicit causal relationships among the events described in the message. The noun phrase analyzer, which must cope with long, complex nominals, relies heavily on the structural and descriptive information in the equipment model. The recovery of causal relations in discourse analysis is based primarily on the simulation capabilities of the model.

This system has been demonstrated publically at four stages during its development: at the DARPA Strategic Computing -- Natural Language Processing meeting in Los Angeles, CA, May 1986; at the Assn. for Computational Linguistics Annl. Meeting in New York, June 1986; at the DARPA Strategic Computing -- Natural Language Processing meeting in Philadelphia, PA, May 1987, and at the Assn. for Computational Linguistics Annl. Meeting in Palo Alto, CA, July 1987.

2.3. Reports and Publications

Proteus Project Memorandum #1: PROTEUS and PUNDIT: Research in Text Understanding. R. Grishman and L. Hirschman, April 1986.

Published in Computational Linguistics **12** (2), 141-145, 1986.

Proteus Project Memorandum #2: Model-based Analysis of Messages about Equipment. R. Grishman, T. Ksiezyk, and N. T. Nhan, April 1986.

Proteus Project Memorandum #3: An Equipment Model and its Role in the Interpretation of Nominal Compounds. T. Ksiezyk and R. Grishman, April 1986.

Proteus Project Memorandum #6: An Equipment Model and its Role in the Interpretation of Noun Phrases. Tomasz Ksiezyk, Ralph Grishman, and John Sterling, January 1987.

Proteus Project Memorandum #6-A: abridged version of PPM #6, April 1987.

Published in Proceedings IJCAI-87, pp. 692-695.

Proteus Project Memorandum #8: Finding Causal and Temporal Relations in Equipment Failure Messages. Leo Joskowicz, Ralph Grishman, and Tomasz Ksiezyk, February 1987.

Proteus Project Memorandum #11: Equipment Simulation for Language Understanding. Tomasz Ksiezyk and Ralph Grishman, September 1987.

3. RAINFORM analysis

In June 1987 we participated in the Message Understanding Conference (MUCK) sponsored by the Naval Ocean Systems Center (NOSC) in San Diego, CA. As part of this conference we were provided by NOSC with a set of RAINFORM sighting messages and, during May 1987 we ported our system to this new domain and performed a syntactic, semantic, and simplified discourse analysis of ten messages specified by NOSC. While at the conference we made (in the course of one morning) further adjustments in order to process an eleventh message given to us by NOSC.

This effort provided an important demonstration of our ability to port the text processing software to a new and quite different domain. It also suggested to us ways in which the software -- particularly the semantic analyzer and the user interface -- could be improved to facilitate future development efforts.

4. Parallel Parsing

4.1. Objective

Parsing remains one of the most time-consuming aspects of natural language processing. The rapid proliferation of small parallel processing systems appears to offer one possible route for substantially speeding up this task. The objective of our experiments was to determine whether, through relatively simple modifications to existing systems, we can obtain substantial speed-ups in parsing.

4.2. Accomplishments

We modified the chart parser of PROTEUS to operate under ZLISP, a parallel LISP developed by Isaac Dimitrovsky as part of the NYU Ultracomputer Project. We ran a series of experiments, both under simulation and on the actual Ultracomputer, varying the complexity of the grammar, the length of the sentence, and the number of processors. Our results were generally encouraging; we obtained typical speed-ups of 5 to 7 with our largest grammar.

4.3. Publication

Proteus Project Memorandum #10: Evaluation of a Parallel Chart Parser. Ralph Grishman and Mahesh Chitrao, September 1987.

To appear in Proc. Second Conf. on Applied Natural Language Processing, Austin, TX, Feb. 1988.

5. Feedback for Semantic Overshoot

5.1. Objective

One of the most difficult tasks in developing a natural language interface involves collecting a complete set of semantic relations and the forms in which they may be expressed. For a large domain and complex application the task may be impossible -- the domain may not be closed or cleanly delineated, and completeness may be beyond reach. We are therefore faced with the problem of handling semantic overshoot: user input which exceeds the semantic model incorporated in the system. This aspect of our research has focussed in providing helpful user feedback in cases of semantic overshoot.

5.2. Accomplishments

Our approach in cases of semantic overshoot has been to identify the closest variants of the user's input which would be acceptable to the system (i.e., would be within the system's semantic

model), and to provide these variants as suggestions to the user. We have conducted a series of experiments using a small question-answering system operating on a domain of student transcripts and course prerequisites. Because of the limitations of the system in terms of vocabulary and syntax, we have found that semantic overshoot is not the primary reason for the rejection of user input; however, in cases of semantic overshoot our system was able, in the majority of cases, to provide appropriate feedback so that the user could reformulate his query.

5.3. Publications

Proteus Project Memorandum #7: Responding to Semantically Ill-formed Input. Ralph Grishman and Ping Peng, January 1987.

Proteus Project Memorandum #7-A: revised version of PPM #7, September 1987.

To appear in Proc. Second Conf. on Applied Natural Language Processing, Austin, TX, Feb. 1988.